Learning low bending and low distortion manifold embeddings

Juliane Braunsmann
University of Münster
Münster, Germany
j.braunsmann@wwu.de

Marko Rajković
University of Bonn
Bonn, Germany

Martin Rumpf
University of Bonn
Bonn, Germany

Benedikt Wirth
University of Münster
Münster, Germany

Abstract

Autoencoders are a widespread tool in machine learning to transform high-dimensional data into a lower-dimensional representation which still exhibits the essential characteristics of the input. The encoder provides an embedding from the input data manifold into a latent space which may then be used for further processing. For instance, learning interpolation on the manifold may be simplified via the new manifold representation in latent space. The efficiency of such further processing heavily depends on the regularity and structure of the embedding. In this article, the embedding into latent space is regularized via a loss function that promotes an as isometric and as flat embedding as possible. The required training data comprises pairs of nearby points on the input manifold together with their local distance and their local Fréchet average. This regularity loss functional even allows to train the encoder on its own. The loss functional is computed via a Monte Carlo integration which is shown to be consistent with a geometric loss functional defined directly on the embedding map. Numerical tests are performed using image data that encodes different data manifolds. The results show that smooth manifold embeddings in latent space are obtained. These embeddings are regular enough such that interpolation between not too distant points on the manifold is well approximated by linear interpolation in latent space.

1. Introduction

A central task in machine learning is to represent objects in high-dimensional data manifolds by points in a lower-dimensional hidden latent space. Methods in this direction can be split into linear and nonlinear approaches. The former include *principal component analysis (PCA)* and *multidimensional scaling (MDS)* [14], examples for the latter are *Isomap* [26], *Local Linear Embedding* [23] and *Hessian Eigenmaps* [5]. Such methods take a collection of high-dimensional data points as input and give a collection of low-dimensional vectors as output. They rely on neighborhood graphs, and a central part of these methods is usually the computation of a spectral embedding and computation of eigenvalues.

A more recent approach to nonlinear dimensionality reduction are a special type of neural networks called *autoencoders*. They consist of an *encoder* and a *decoder*. The encoder maps from the high-dimensional ambient space of the data manifold to a low-dimensional Euclidean space, called *latent space*. The decoder maps from latent space back to the ambient space of the data manifold and tries to reproduce the original input data. In the training phase, the encoder and decoder mapping are determined via the minimization of a loss functional. The image of a smooth data manifold via a smooth encoder map is a smooth submanifold in the Euclidean latent space. The assumption that the observed high-dimensional data actually forms a low-dimensional manifold – the image of the latent manifold under the decoder map – is called the *manifold hypothesis*.

In *deep manifold learning* one aims at recovering a simpler, low-dimensional latent manifold representation of the data manifold from the observed data via the minimization of a loss functional. In the first place, this loss functional measures the *reconstruction loss* by comparing the input data with its image under the composition of encoder and decoder mapping. Since neural networks can be arbitrarily complex, though, a focus solely on the reconstruction loss risks that the autoencoder simply "learns by heart". Different strategies have been investigated in addition to the reconstruction loss which favor smoothness of the encoder and decoder mapping and thus regularity of the latent manifold. Among them are methods which promote sparsity [20], contractive autoencoders [22], or denoising autoencoders [28]. In [22] the loss functional penalizes the norm of the Jacobian of the encoder to achieve robustness to small changes of the input around the training samples.

Data manifolds often come with a metric encoding the cost of local variations on the manifold. For sufficiently regular data manifolds it is shown in [25] how to transfer this metric to the latent manifold and thereby make it *Riemannian*. This allows to compute shortest paths, exponen-

tial maps, and parallel transport in latent space, which the decoder can push forward to the data manifold.

Autoencoders also offer the possibility to interpolate between data points by interpolating linearly in the latent space. In [4] an *adversarial regularizer* was proposed to ensure visually realistic interpolations in latent space. The adversarial regularizer tries to make the decoding of interpolations in latent space indistinguishable from real data points. Recently, a generalized definition of interpolation via the training of a discriminator was proposed in [16] which allows to check that the interpolated point belongs to the original dataset. In this approach an additional smoothness loss is used based on differentiation along interpolation paths in latent space. While the method in [4] relies solely on an adversarial network to discriminate between real data and interpolations, the approach in [16] also suggests to include ground truth interpolation data.

A major deficit of autoencoders is that they frequently fail to reproduce the statistical input data distribution in the latent space. In [12] *isometric*, *i.e.*, length-preserving encoder maps are used to more accurately push forward distributions from input to latent space. To this end a loss function based on Shannon-Rate-Distortion theory is proposed.

The loss functional from [3] promotes isometry of the decoder map (by penalizing deviation from a non-orthogonal Jacobian matrix) and that the encoder is a pseudo-inverse of the decoder (by enforcing the Jacobians of de- and encoder to be transposes of each other). In [18] isometric embeddings in latent space are learned to obtain standardized data coordinates from scientific measurements. The authors approximate the Jacobian through normally distributed sampling around each data point (so-called *bursts*), and an objective functional measures the lack of orthogonality of the Jacobian via the deviation of the local covariance of bursts from the identity. In [24], with the goal of producing a globally isometric encoder map, a loss functional is proposed which measures the difference between distances in the pushforward metric and distances in latent space. The training of encoders in [17] is based on a loss functional which compares Euclidean distances in latent space with geodesic distances on the input manifold.

Our contribution. This paper investigates a loss functional for the geometric regularization of the latent manifold. Such a loss requires some geometric data of the input manifold in the training. We follow here a minimalistic approach involving data on distances and averages on the input manifold. We provide as testbed examples for which this data is explicitly known and easy to compute.

- We propose to complement a *discrete loss* functional that promotes *isometric embeddings* in latent space with a discrete *bending loss* functional which prefers as flat as possible embeddings; this combination optionally permits to train the encoder on its own.
- To train a corresponding autoencoder network we consider training data consisting of *triplets* of two input data points and their *Fréchet average* together with the distance between the two data points.
- Unlike other isometry promoting approaches we do not approximate the *Jacobian of the encoder* (via backpropagation or complete correlation of all nearest neighbors) but use simple *Monte Carlo point sampling*.
- Matching the theory of isometric maps, our numerical experiments confirm that the bending loss significantly increases *smoothness* of the resulting latent space manifold over a pure isometry loss. Also we find that the decoder maps *linear interpolation in latent space* to reasonable interpolations on the data manifold.
- We demonstrate that our discrete loss functional is consistent with a *well-posed, continuous limit functional* on encoder maps from a smooth Riemannian data manifold into latent space.

The paper is structured as follows. Section 2 derives the new regularization loss and discusses its Monte Carlo limit for dense sampling of point pairs. The limit functional of this Monte Carlo limit for vanishing distance between the point pairs is introduced in section 3, and we prove existence of minimizers. As a proof of concept, we train autoencoders on three image datasets representing a priori known data manifolds. Section 4 describes the autoencoder set up, the experimental datasets, and the numerical results.

2. A low bending and low distortion regularization for encoders

Consider a smooth compact m-dimensional Riemannian manifold (M, g) possibly with boundary, of which we have samples available. To avoid technical details the theory will be presented only for input manifolds without boundary. We assume that M is embedded in some very high-dimensional space $\mathbb{R}^n$, for instance the space of images with n pixels. The aim is now to compute an embedding of M into Euclidean space $\mathbb{R}^l$ (called the *latent space*), where we think of the dimension l as being only moderately larger than m (for instance $l = 2m$ so that the existence of a smooth embedding is guaranteed by Whitney's embedding theorem). Such a representation is often learned from samples by training an autoencoder, which is a pair of maps

$$\phi : M \to \mathbb{R}^l,\ \psi : \mathbb{R}^l \to \mathbb{R}^n \text{ with } \psi(\phi(x)) \approx x \text{ for all } x \in M.$$

The image $\phi(M)$ of M is called *latent manifold*. The autoencoder functions ϕ and ψ are implemented as deep neural networks. An appropriate structure and regularity of the embedding ϕ into latent space is known to aid downstream tasks such as classification, Riemannian interpolation and

extrapolation, clustering or anomaly detection. For this reason we aim for a natural, geometrically inspired regularizing loss function for the encoder ϕ.

From the viewpoint of downstream processing, the nicest embedding $\phi : M \to \mathbb{R}^l$ would of course be an isometric embedding into an affine subspace of $\mathbb{R}^l$. This would identify M as isometric to flat Euclidean space so that any downstream processing on M can be performed with the simplicity and efficiency immanent to Euclidean space. In particular, the most basic operations of computing distances and interpolations would become trivial. Of course, such an embedding is usually prevented by the intrinsic or the global geometry of M, nevertheless one may try to get as close as possible to an **I**sometric and **F**lat embedding, at least locally. Hence we suggest the following two simple objectives for any two not too distant points $x, y \in M$:

(**I**) The intrinsic Riemannian distance between x and y in M should differ as little as possible from the Euclidean distance between the latent codes $\phi(x)$ and $\phi(y)$.

(**F**) A (weighted) average between x and y in M should deviate as little as possible from the (weighted) Euclidean average between $\phi(x)$ and $\phi(y)$.

If the first objective is only applied to infinitesimally close points $x, y \in M$, it is nothing else than asking for an *isometric embedding*, as extensively pursued in the literature (*cf*. section 1). It ensures that ϕ embeds M in latent space with low distortion. However, asking for isometry alone is highly questionable from the mathematical point of view since the family of isometric embeddings is very large and contains quite irregular elements (Nash–Kuiper embeddings are in general only Hölder differentiable). Therefore, in (**I**) and (**F**) we go beyond this infinitesimal isometry viewpoint:

- The isometry objective (**I**) asks that the *intrinsic* distances between x and y in M are approximated by the *extrinsic* distances between $\phi(x), \phi(y) \in \mathbb{R}^l$ in latent space (rather than the intrinsic distances in the latent manifold $\phi(M)$, which would define an isometric embedding).
- The flatness/bending objective (**F**) enforces some second order low bending regularity or flatness on ϕ by requiring that the *geodesic interpolation* between x and y in M is well approximated by extrinsic *linear interpolation* in the latent space $\mathbb{R}^l$.

A low bending and low distortion loss. Denote the geodesic distance between two points $x, y \in M$ by $d_M(x, y)$ and their geodesic average by $\mathrm{av}_M(x, y)$. As input data to the training or optimization of the encoder ϕ we consider a sample $\mathcal{S}_\epsilon \subset \{(x, y) \in M \times M \,|\, d_M(x, y) \le \epsilon\}$ of pairs $(x, y) \in M \times M$ of nearby points together with $d_M(x, y)$ and $\mathrm{av}_M(x, y)$. For a sufficiently small fixed locality radius $\epsilon > 0$, the unique existence of the geodesic average is ensured. Our proposed loss function to regularize the encoder ϕ then reads

$$E^{\mathcal{S}_\epsilon}(\phi) = \frac{1}{|\mathcal{S}_\epsilon|} \sum_{(x,y)\in\mathcal{S}_\epsilon} \left(\gamma(\partial_{(x,y)}\phi) + \lambda\, |\partial^2_{(x,y)}\phi|^2\right)$$

with first order difference quotient $\partial_{(x,y)}\phi = \frac{\phi(y)-\phi(x)}{d_M(x,y)}$ and second order difference quotient $\partial^2_{(x,y)}\phi = 8\frac{\mathrm{av}_{\mathbb{R}^l}(\phi(x),\phi(y))-\phi(\mathrm{av}_M(x,y))}{d_M(x,y)^2}$, where $\gamma(s) = |s|^2 + |s|^{-2} - 2$, $\mathrm{av}_{\mathbb{R}^l}(a, b) = (a + b)/2$ denotes the linear average in $\mathbb{R}^l$, and $\lambda > 0$. Note that the first term in $E^{\mathcal{S}_\epsilon}$ has a strict minimum for $|\partial_{(x,y)}\phi| = 1$. This term thus promotes $|\phi(x) - \phi(y)| = d_M(x, y)$ and thus low distortion and approximate isometry. The second term in $E^{\mathcal{S}_\epsilon}$ penalizes the deviation of intrinsic averages on $\phi(M)$ from extrinsic ones in $\mathbb{R}^l$. Note that this does not only penalize bending or any extrinsic curvature of $\phi(M)$ in $\mathbb{R}^l$, but in addition it also penalizes deviation of the inplane parameterization of $\phi(M)$ from a linear one (*cf*. the corresponding remark for $\mathrm{Hess}\,\phi$ in the next section). Examples of how to compute av_M and d_M for image input data include the corresponding methods from the theories of LDDMM [29], metamorphosis [27], or optimal transport [19]. In our testbed we purposely used low-dimensional manifolds where av_M and d_M are explicitly known.

The Monte Carlo limit for dense sampling. Assuming that $\mathcal{S}_\epsilon$ is drawn uniformly from $M \times M$ (subject to the locality condition), our loss function $E^{\mathcal{S}_\epsilon}$ is up to $\mathcal{O}(\epsilon)$ the Monte Carlo integration of the energy

$$\mathcal{E}^\epsilon(\phi) = \fint_M \fint_{B^M_\epsilon(x)} \gamma(\partial_{(x,y)}\phi) + \lambda\, |\partial^2_{(x,y)}\phi|^2 \mathrm{d}V_g(y)\mathrm{d}V_g(x), \quad (1)$$

where $B^M_\epsilon(x)$ denotes the geodesic ϵ-ball in M, centered at x, and where $\fint \ldots \mathrm{d}V_g$ denotes the mean with respect to the Riemann–Lebesgue volume measure on M (the index g indicates the Riemannian metric). As for the discrete functional $E^{\mathcal{S}_\epsilon}$, the energy $\mathcal{E}^\epsilon$ penalizes deviation from isometry and from intrinsically and extrinsically flat embeddings.

The energy $\mathcal{E}^\epsilon$ is rigid motion invariant by construction, *i.e*., composition of ϕ with a rigid motion does not change the energy. However, even apart from this invariance one cannot expect uniqueness of minimizers due to the nonconvexity of the first integrand. This is unavoidable when promoting isometries, *e.g*., if M is represented in latent space as a half sphere $\phi(M)$, then an equivalent embedding would be obtained for the half sphere flipped inside out. Whenever M is intrinsically flat and homeomorphic to the m-disc (or at least globally compatible with an embedding into an m-dimensional Euclidean space), there is a unique minimizer of $\mathcal{E}^\epsilon$, though, as stated in the following proposition.

Proposition 1 (unique embedding of intrinsically flat discs)**.** *If M is the flat m-disc D^m, the unique minimizer (up to rigid motion) of $\mathcal{E}^\epsilon$ is $\phi : M \ni x \mapsto (x, 0, \dots, 0) \in \mathbb{R}^l$.*

Proof. It is straightforward to check $\mathcal{E}^\epsilon \geq 0$ as well as $\mathcal{E}^\epsilon(\phi) = 0$ so that ϕ is a global minimizer. The uniqueness up to rigid motion then follows from the fact that fixing ϕ at $m+1$ points $x_0, \dots, x_m$ uniquely determines ϕ at all points x within the convex hull of $x_0, \dots, x_m$ since such x can be represented as (limit of) iterated averages of $x_0, \dots, x_m$ so that $\phi(x)$ must be the (limit of the) corresponding iterated averages of $\phi(x_0), \dots, \phi(x_m)$ (that one may take limits of ϕ follows from the condition $\partial_{(x,y)}\phi = 1$ for all close enough x, y). However, fixing $m+1$ points with prescribed distances just fixes a rigid motion. □

Note that this property of recovering flat embeddings may be quite relevant in applications as generative image manifolds were noticed in [25] to have almost no curvature.

Affinely invariant loss functions. In several applications one may already be content with a nice embedding ϕ that is specified only up to an affine transformation (rather than a rigid motion). Indeed, one may want to abandon isometry in favor of improving the approximation of geodesic averages by linear averages. This raises the question whether one can replace the integrand in (1) by some function $f(\partial_{(x,y)}\phi, \partial^2_{(x,y)}\phi)$ which is invariant under left composition of ϕ with invertible affine (and not just rigid) transformations. Yet, if $\partial_{(x,y)}\phi$ and $\partial^2_{(x,y)}\phi$ are non-parallel there is always an affine transform that maps $\partial_{(x,y)}\phi$ onto $(1, 0, \cdots)$ and $\partial^2_{(x,y)}\phi$ onto $(0, 1, 0, \cdots)$ so that necessarily f is of the form $f(a, b) \equiv c$ for a, b non-parallel and $f(a, b) = h(s)$ for $b = sa$ with some constant c and function $h : \mathbb{R} \to \mathbb{R}$. Thus, the regularizing properties would be lost. An alternative could be to just penalize $|\partial^2_{(x,y)}\phi|/|\partial_{(x,y)}\phi|$. Though not affinely invariant, it still encodes that geodesics should be close to linear interpolation without any competing isometry constraints ($|\partial_{(x,y)}\phi|$ in the denominator is needed for scale invariance and prevents a collapse to $\phi(M) = 0$). Our loss $\mathcal{E}^\epsilon$ controls this flatness measure due to

$$2\sqrt{\lambda}|\partial^2_{(x,y)}\phi|/|\partial_{(x,y)}\phi| \leq |\partial_{(x,y)}\phi|^{-2} + \lambda\,|\partial^2_{(x,y)}\phi|^2\,.$$

Replacing $\mathcal{E}^\epsilon(\phi)$ with $⨍_M ⨍_{B^M_\epsilon(x)} |\partial^2_{(x,y)}\phi|/|\partial_{(x,y)}\phi| \mathrm{d}y\,\mathrm{d}x$ would be infeasible, though: a straightforward calculation shows that a minimizing sequence of embeddings of the cylinder $M = S^1 \times [0, 1]$ into $\mathbb{R}^l$ would be cylinders of vanishing radius and diverging length.

3. The limit of vanishing locality radius

An appropriate structure and regularity of an embedding $\phi : M \to \mathbb{R}^l$ can also be promoted by a purely local functional. Below we present a natural candidate and identify it as a consistent limit of $\mathcal{E}^\epsilon$ when $\epsilon \to 0$ under smoothness assumptions on the embedding.

A purely local low bending and low distortion loss. The *Riemannian gradient (Jacobian)* $\operatorname{grad}\phi(x) \in (T_xM)^l$ of ϕ is defined (denoting standard differentiation of smooth extensions onto $\mathbb{R}^n$ by D) via the identity

$$\mathrm{D}\phi_j(x)(v) = \tfrac{\mathrm{d}}{\mathrm{d}t}(\phi_j \circ \exp_x)(tv)|_{t=0} = g(\operatorname{grad}\phi_j(x), v)$$

for all $v \in T_xM$, where $\exp_x : T_xM \to M$ denotes the Riemannian exponential map in x. An isometric embedding $\phi : M \to \mathbb{R}^l$ is characterized by $\operatorname{grad}\phi(x)$ being orthogonal in any point $x \in M$. Thus, deviation from an isometric embedding manifests as non-*unit singular values of* $\operatorname{grad}\phi(x)$. Similarly, extrinsic bending of the embedding $\phi(M)$ manifests as a non-*vanishing Riemannian Hessian* $\operatorname{Hess}\phi$ of ϕ, where the Riemannian Hessian at x is the linear operator $\operatorname{Hess}\phi(x) : T_xM \to (T_xM)^l$, $\operatorname{Hess}\phi(x)(v) = (\nabla_v \operatorname{grad}\phi_j)_{j=1,\dots,l}$ for ∇ the Levi-Civita connection and ∇_v the covariant derivative in direction v [1, Chp. 5]. The associated quadratic form $\mathrm{D}^2\phi(x) : T_xM \times T_xM \to \mathbb{R}^l$ is

$$\begin{aligned}\mathrm{D}^2\phi_j(x)(v, v) &= \tfrac{\mathrm{d}^2}{\mathrm{d}^2 t}(\phi_j \circ \exp_x(tv))|_{t=0} \\ &= g(\operatorname{Hess}\phi_j(x)(v), v),\end{aligned}$$

where the Riemannian metric g on $(T_xM)^l \times T_xM$ is applied componentwise, *i.e.*, we use the notation $g(A, v) = (g(A_j, v))_{j=1,\dots,l}$ for a matrix whose rows A_j are tangent vectors. A natural loss function to promote low distortion and low bending embeddings thus reads

$$\mathcal{E}(\phi) = ⨍_M \Gamma(\operatorname{grad}\phi(x)) + \tfrac{\lambda}{2}\|\operatorname{Hess}\phi(x)\|_F^2 \mathrm{d}V_g(x), \quad (2)$$

where $\|A\|_F^2 = \operatorname{tr}(A^*A)$ is the Frobenius norm of an operator A and $B \mapsto \Gamma(B)$ is an *admissible* function, *i.e.*, a nonnegative function depending only on the singular values of B and being zero if and only if all of them are one.

Let us remark that the orthogonal projection of $\operatorname{Hess}\phi(x)$ onto the normal bundle $[T_{\phi(x)}\phi(M)]^\perp$ of $\phi(M)$ is the second fundamental form of $\phi(M)$ pulled back onto T_xM. If $m = 2$ and $l = 3$, this is also known as the *Weingarten map* or *shape operator relative to* T_xM. In addition to penalizing this second fundamental form, which indicates extrinsic bending of $\phi(M)$, our functional $\mathcal{E}$ also penalizes the tangential components of $\operatorname{Hess}\phi$.

Reformulation with directional derivatives. The above energy can be rewritten in terms of averages. To this end, let S^{m-1} denote the unit sphere in T_xM (the base point x will be clear from the context), and let $\mathcal{H}^{m-1}$ denote the $(m-1)$-dimensional Hausdorff measure in T_xM.

Proposition 2 (double integral representation of $\mathcal{E}$)**.** *The choice* $\Gamma : (T_xM)^l \to [0, \infty]$,

$$\Gamma(B) = ⨍_{S^{m-1}} \gamma(g(B, v))\,\mathrm{d}\mathcal{H}^{m-1}(v)$$

is admissible in the above sense and leads to the representation

$$\mathcal{E}(\phi) = \fint_M \fint_{S^{m-1}} \gamma(\mathrm{D}\phi(x)(v)) + \lambda|\,\mathrm{D}^2\phi(x)(v,v)|^2 \mathrm{d}\mathcal{H}^{m-1}(v)\mathrm{d}V_g(x).$$

Proof. Let B have singular values $\sigma_1,\dots,\sigma_m$ and left and right singular vectors $w_1,\dots,w_m \in \mathbb{R}^l$, $v_1,\dots,v_m \in T_xM$, then $|g(B,v)|^2 = |\sum_{i=1}^m \sigma_i g(v_i,v)w_i|^2 = \sum_{i=1}^m \sigma_i^2 g(v_i,v)^2$. Inserting this expression into Γ, one sees that the integral makes the expression independent of the orthonormal frame $v_1,\dots,v_m$ so that $\Gamma(B)$ indeed only depends on the singular values of B. Furthermore, Γ is nonnegative since its integrand is, and it is zero if and only if $|g(B,v)| = 1$ for all $v \in S^{m-1}$ in T_xM, which by the above is equivalent to all singular values being one. Analogously one shows that

$$\fint_{S^{m-1}} \sum_{j=1}^l |\,\mathrm{D}^2\phi_j(x)(v,v)|^2 \mathrm{d}\mathcal{H}^{m-1}(v) = \sum_{j=1}^l \sum_{i=1}^m (\sigma_i^j)^2 \fint_{S^{m-1}} g(v,e)^2 \mathrm{d}\mathcal{H}^{m-1}(v) = \tfrac{1}{2}\|\operatorname{Hess}\phi(x)\|_F^2$$

for the eigenvalues σ_i^j of $\operatorname{Hess}\phi_j(x)$ and some arbitrary $e \in S^{m-1}$ using that the integral inside the sum is $\frac{1}{2}$. □

Identification as limit for vanishing locality radius. It turns out that for $\epsilon \to 0$ our loss $\mathcal{E}^\epsilon$ approximates $\mathcal{E}$, which thus gives a simple interpretation of $\mathcal{E}^\epsilon$ in terms of first and second order derivatives of ϕ.

Theorem 1 (limit energy for vanishing ϵ)**.** *$\mathcal{E}^\epsilon$ is a consistent approximation of $\mathcal{E}$ in the sense $\mathcal{E}^\epsilon(\phi) = \mathcal{E}(\phi) + \mathcal{O}(\epsilon\|\phi\|_{C^3})$.*

Proof. For ϵ small enough, the Riemannian exponential $\exp_x$ defines a diffeomorphism between the ϵ-ball $B_\epsilon(0) \subset T_xM \cong \mathbb{R}^m$ and $B_\epsilon^M(x)$ with inverse denoted by $\log_x$. For any measurable function $f_x : M \to \mathbb{R}$ we then have

$$\fint_{B_\epsilon^M(x)} f_x(y)\mathrm{d}V_g(y) = \fint_{B_\epsilon(0)} f_x(\exp_x w)\mathrm{d}(\log_x^* V_g)(w),$$

where $\log_x^* V_g$ is the pushforward measure of V_g under $\log_x$. The Lebesgue density of $\log_x^* V_g$ at $w \in B_\epsilon(0)$ is known to have the expansion $1 + \mathcal{O}(|w|^2)$ (the constant depends on the Ricci curvature, *cf.* [2]). This can be used together with the transformation formula to get

$$\begin{aligned}&\fint_{B_\epsilon(0)} f_x(\exp_x w)\mathrm{d}(\log_x^* V_g)(w)\\ &= \fint_{B_1(0)} f_x(\exp_x(\epsilon w))(1+\mathcal{O}(\epsilon^2))\,\mathrm{d}w\\ &= \fint_{S^{m-1}} \int_0^1 f_x(\exp_x(\epsilon r v))(1+\mathcal{O}(\epsilon^2)) m r^{m-1}\mathrm{d}r\mathrm{d}\mathcal{H}^{m-1}(v).\end{aligned}$$

Now we consider the cases $f_x(y) = \gamma(\partial_{(x,y)}\phi)$ as well as $f_x(y) = |\partial^2_{(x,y)}\phi|^2$. Letting $y = \exp_x(\epsilon r v)$ and abbreviating $\theta(t) = \phi(\exp_x(tv))$, Taylor expansion yields

$$\partial_{(x,y)}\phi = \frac{\theta(\epsilon r)-\theta(0)}{r\epsilon} = \theta'(0) + \mathcal{O}(r\epsilon) \quad \text{and}$$
$$\partial^2_{(x,y)}\phi = 8\frac{\frac{1}{2}(\theta(0)+\theta(\epsilon r)) - \theta(\frac{\epsilon r}{2})}{r^2\epsilon^2} = \theta''(0) + \mathcal{O}(r\epsilon).$$

Now by the definition of gradient and Hessian we have $\theta'(0) = \mathrm{D}\phi(x)(v)$ and $\theta''(0) = \mathrm{D}^2\phi(x)(v,v)$. The proof is concluded by inserting these estimates in f_x and noting that the constants of all error terms in ϵ depend on the manifold M and on (at most) third derivatives of ϕ. □

Existence of optimal geometric embeddings. Let us now establish the existence of minimizers to $\mathcal{E}$. First, we observe that the energy $\mathcal{E}$ is well-defined on all of $H^2(M)$, where the Sobolev space $H^2(M)$ is defined as the closure of all smooth functions under the norm $\|\phi\|_{H^2(M)}$ with

$$\|\phi\|^2_{H^2(M)} = \sum_{j=1}^m \int_M |\phi_j|^2 + g(\operatorname{grad}\phi_j, \operatorname{grad}\phi_j) + g(\operatorname{Hess}\phi_j, \operatorname{Hess}\phi_j)\,\mathrm{d}V_g. \quad (3)$$

For further details on Sobolev spaces on (compact) manifolds we refer to [9]. Due to the rigid motion invariance of $\mathcal{E}$ we may without loss of generality restrict $\mathcal{E}$ to the subspace $\dot{H}^2(M)$ of H^2-functions with zero average.

Theorem 2 (existence of a minimizer)**.** *Let M be smooth, compact. If there exists $\phi \in \dot{H}^2(M)$ with $\mathcal{E}(\phi) < \infty$, then $\mathcal{E}$ has a minimizer in $\dot{H}^2(M)$.*

If $l \geq 2m$ the condition is always fulfilled since by Whitney's embedding theorem there exists a smooth embedding which, due to the compactness of M, may be chosen such that it has finite energy.

Proof. We apply the direct method in the calculus of variations. By our assumption there exists a minimizing sequence $(\phi^k)_{k=1,2,\dots} \subset \dot{H}^2(M)$, which we suppose to converge monotonically to $\inf\mathcal{E} < \infty$. Since $\Gamma(\operatorname{grad}\phi) \geq C|g(\operatorname{grad}\phi,\operatorname{grad}\phi)| - 2$ and $\|\operatorname{Hess}\phi\|_F^2 \geq C|g(\operatorname{Hess}\phi,\operatorname{Hess}\phi)|$ for some constant $C > 0$, the second and third summand of (3) are uniformly bounded for all ϕ^k. By Poincaré's inequality this implies uniform boundedness of ϕ^k in $\dot{H}^2(M)$. By reflexivity of $\dot{H}^2(M)$, there exists a weakly convergent subsequence (still indexed by k) with limit ϕ in $\dot{H}^2(M)$. Convexity of the map $A \to \|A\|_F^2$ then implies $\liminf_{k\to\infty} \fint_M \|\operatorname{Hess}\phi^k\|_F^2 \mathrm{d}V_g \geq \fint_M \|\operatorname{Hess}\phi\|_F^2 \mathrm{d}V_g$. Furthermore, by Rellich embedding, $\operatorname{grad}\phi^k$ already converges strongly to $\operatorname{grad}\phi$ in $L^2(M)$ and up to selection of another subsequence even pointwise almost everywhere. Fatou's lemma then implies

$\int_M \Gamma(\operatorname{grad}\phi)\mathrm{d}V_g \le \liminf_{k\to\infty} \int_M \Gamma(\operatorname{grad}\phi^k)\mathrm{d}V_g$. Thus, we obtain lower semi-continuity of the energy, *i.e.*, $\mathcal{E}(\phi) \le \liminf_{k\to\infty} \mathcal{E}(\phi^k) = \inf \mathcal{E}$, which establishes the claim. □

Just as for $\mathcal{E}^\epsilon$, the minimizer (modulo a rigid motion) is in general not unique due to the isometry promoting term.

4. Numerical experiments

In what follows, similarly to [6], we consider image data that implicitly represent three different manifolds:

(G) images of anisotropic Gaussians which are rotated, scaled and translated, representing a cylinder $S^1 \times [a,b] \times [c,d]^2$,

(S) shadows of a sundial with the sun or light source shining from all possible directions, representing the upper hemisphere $S^2 \cap \{x_3 \ge 0\}$ (*cf.* [16]),

(R) orthogonal projections of a rotated 3D object, representing $SO(3)$.

Datasets. We consider an image resolution of 64×64. The images are generated as follows.

(G) Anisotropic Gaussians. We consider rotations, scalings, and translations of a cut off Gaussian of fixed aspect ratio, with parameters $(\alpha, s, x) \in M = S^1 \times [a,b] \times [c,d]^2$ with distance $d((\alpha,s,x_1,x_2),(\alpha',s',x_1',x_2'))^2 = d_{S^1}(\alpha,\alpha')^2 + |s-s'|^2 + |x_1-x_1'|^2 + |x_2-x_2'|^2$, where d_{S^1} is the geodesic distance on S^1. The data is similar to the DSprites dataset in [15].

(S) Sundials. Inspired by [16], we generate images parametrized by the upper hemisphere $M = S^2 \cap \{x_3 \ge 0\}$ by casting a shadow of a vertical rod on a plane. Contrary to [16] we do not render these images with a 3D engine, but instead simply approximate the shadows by Gaussians (*cf.* fig. 1): a point $x \in M$ is first mapped onto the plane by drawing the line through x and the rod tip, intersecting the plane at some $y \in \mathbb{R}^2$. We then use a Gaussian function with variance $|y|$ in direction y and a fixed small variance in the orthogonal direction, centered at $y/2$. As distance on M we use the geodesic distance on S^2, $d(x,x') = \arccos(x^T x')$.

(R) Rotated 3D objects. We generate images by rotating a camera pointing at a three-dimensional object, Spot the cow. We use `Pytorch3D` [21] to render the images during training. As distance on $M = SO(3)$ we use the geodesic distance computed via quaternions as $d(q_1,q_2) = \arccos(|q_1 \cdot q_2|)$ [11].

Autoencoder architecture. The used architecture is as in [4], however, we used larger input images and a smaller latent dimensionality. The encoder consists of a first layer of 1×1 convolutions with 16 output channels, followed by blocks consisting of two consecutive 3×3 convolutions with unit stride, zero padded such that the input and output width are equal, and 2×2 average pooling. Each of the convolutional layers in the block is followed by a leaky ReLU nonlinearity with slope -0.01. The first convolution in each block doubles the output channels. The final convolution is not followed by a nonlinearity and has only 1 (4 for (R)) output channel(s). The number of convolutional blocks determines the size of the latent code: we used 4 blocks with input images of size 64×64 ($64 \times 64 \times 3$ for (R)), resulting in a latent code of size 16 (64 for (R)). The decoder also consists of consecutive blocks of two 3×3 convolutions with leaky ReLU nonlinearities, where each block is followed by 2×2 nearest neighbor upsampling. The final convolutional layer is again not followed by a nonlinearity and has 1 output channel. We use Kaiming initialization [8], *i.e.*, all convolutional weights are initialized as zero-mean Gaussian random variables with standard deviation $\sqrt{2}/\sqrt{\texttt{fan\_in}(1+0.01^2)}$ for `fan_in` the layer input dimension, and all biases are initialized with zeros. For training, we use the Adam optimizer [13] with learning rate 0.0001 and default values for β_1, β_2 and ε. The training data are triplets of images plus a distance value (*cf.* fig. 1 top), $(x, y, \mathrm{av}_M(x,y), d_M(x,y))$, with $x, y \in \mathcal{S}_\epsilon \subset M$, $\mathrm{av}_M(x,y)$ the geodesic average of x, y in M and $d_M(x,y)$ their geodesic distance. This input allows to compute the ingredients $\partial_{(x,y)}\phi$ and $\partial^2_{(x,y)}\phi$ of the loss functional $E^{\mathcal{S}_\epsilon}(\phi)$ defined in 2.

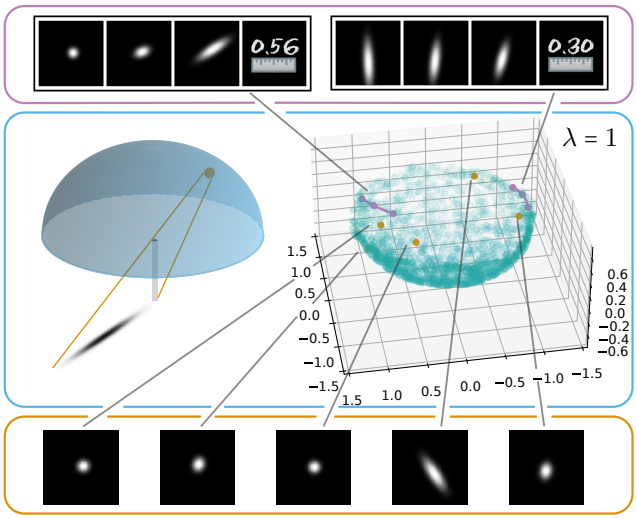

Figure 1. Results for the sundial dataset (S). The top box shows selected training data (image pairs, geodesic average, distance). The middle box shows a sketch of the sundial configuration and the latent manifold $\phi(M)$ projected into $\mathbb{R}^3$ via PCA. The bottom box shows decoder outputs for the orange points in latent space.

Smooth embedding and reliable reconstruction. Figure 1 summarizes our approach and its result at one glance for dataset (S); the top box shows examples of training triplets plus distances, the middle box visualizes the obtained manifold embedding $\phi(M)$, and the bottom box displays reconstructions of input images by the full autoencoder. Fig-

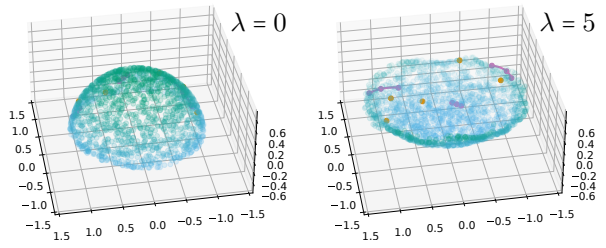

Figure 2. Latent manifold $\phi(M)$ for sundial dataset (S), obtained for different values of flatness weight λ (colored points as in fig. 1). Due to rigid motion and symmetry invariance of $E^{\mathcal{S}_\epsilon}$, the latent manifold appears in orientations different from fig. 1.

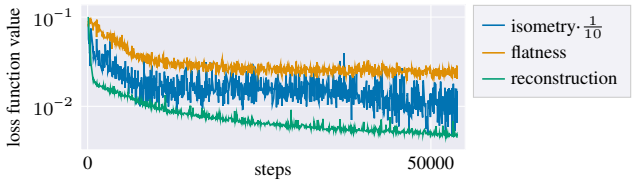

Figure 3. Temporal evolution of the three loss components for dataset (R) of rotated objects (logarithmic y-axis, value of isometry loss scaled down by factor 10). Per optimization step a batch of 128 images is processed.

ures 2, 5 and 6 show further obtained manifold embeddings in latent space for different loss weights and for the datasets (R) and (G). In all cases we observe smooth embeddings that neatly reproduce the geometry and topology of the underlying manifold M. For visualizing the manifold embeddings $\phi(M)$ we simply perform a principal component analysis (PCA) in latent space and display the resulting top three dimensions (in fig. 5 a second set of principal components is shown in addition). To illustrate that our approach allows separate training of encoder and decoder, for fig. 1 we first trained the encoder map ϕ on its own by minimizing $E^{\mathcal{S}_\epsilon}(\phi)$ and subsequently trained the decoder map ψ by minimizing, for fixed ϕ, the *reconstruction loss*

$$R(\phi,\psi) = \frac{1}{|\mathcal{S}_\epsilon|}\sum_{(x,y)\in\mathcal{S}_\epsilon} \|\psi(\phi(x))-x\|_{L^2}^2 + \|\psi(\phi(y))-y\|_{L^2}^2$$

with $\|\cdot\|_{L^2}$ the L^2-norm on images. For the other datasets we train en- and decoder simultaneously by jointly minimizing $E^{\mathcal{S}_\epsilon}(\phi) + \kappa R(\phi,\psi)$ for ϕ and ψ (where the weight $\kappa > 0$ is not expected to have much influence since ψ will try to minimize $R(\phi,\psi)$ anyhow). We use $\epsilon = \frac{\pi}{2}$ for dataset (S), $\epsilon = \frac{\pi}{4}$ for (R). For (G), for simplicity we used a slightly different sample $\hat{\mathcal{S}}_\epsilon \subset \{(y,y') \in M\times M \,|\, d_{S^1}(\alpha,\alpha') \le \epsilon\}$ with $\epsilon = \frac{\pi}{2}$. During training the isometry, flatness, and reconstruction parts of the loss function are all observed to decrease continuously and monotonically (up to the usual stochastic variations) as shown in fig. 3 for dataset (R),

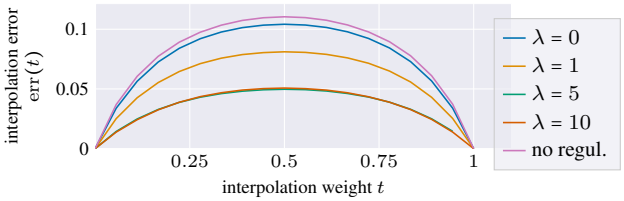

Figure 4. Average error of linear interpolation in latent space for dataset (R) and different flatness weights λ.

where the full autoencoder is trained simultaneously. Resulting reconstructions $\psi(\phi(x))$ for random $x \in M$ are exemplarily shown in figs. 1 and 6 and are of good quality. Figure 2 illustrates the effect of the flatness term: for zero bending penalization λ one obtains a standard isometric embedding of S^2, while it gets flattened for higher values of λ (fig. 1 shows an intermediate λ). Note that extrinsic bending is obviously reduced this way, while nonlinear inplane distortion is increased (the normal component of the Hessian apparently outweighs the inplane component).

Linear interpolation in latent space. Figure 5 illustrates for dataset (R) that the bending term of our loss functional strongly improves the usefulness of linear interpolation in latent space across moderate distances. While for $\lambda = 0$ the decoder output of linear interpolations in latent space does not at all reproduce continuously rotating objects, it clearly does for bending weight $\lambda = 10$. Let us emphasize that the decoder was in no way regularized in these experiments, in particular it was not trained on any points obtained via linear interpolation of codes in latent space! Still the rotated cow is cleanly visible (in contrast to the case $\lambda = 0$), having undergone merely some minor smoothing. Training the decoder additionally on linear interpolants in latent space will naturally improve the results. We purposely abstained from this extra regularization since it would allow the decoder to compensate the deficiencies visible for $\lambda = 0$ so that the regularizing properties of our encoder loss functional $E^{\mathcal{S}_\epsilon}$ would be obscured. For $\lambda = 10$ the decoder has to compensate much less and is therefore expected to be more robustly trainable. We also quantitatively evaluated the quality of linear interpolation in latent space by measuring the L^2-error to the ground truth, geodesic interpolation. We calculate this on a test sample set $\mathcal{S}'_\epsilon$ as

$$\mathrm{err}(t)^2 = \frac{1}{|\mathcal{S}'_\epsilon|}\sum_{(x,y)\in\mathcal{S}'_\epsilon} \mathrm{err_i}(x,y;t)^2 - \mathrm{err_b}(x,y;t)^2 \text{ for}$$

$$\mathrm{err_i}(x,y;t) = \|\,\mathrm{av}_M(x,y;t) - \psi(\mathrm{av}_{\mathbb{R}^l}(\phi(x),\phi(y);t))\|_{L^2},$$

$$\mathrm{err_b}(x,y;t) = \|\,\mathrm{av}_M(x,y;t) - \psi(\phi(\mathrm{av}_M(x,y;t)))\|_{L^2},$$

where $\mathrm{av}_M(x,y;t)$ is the weighted geodesic average of $x,y \in M$ with weights $1-t, t$ and $\mathrm{av}_{\mathbb{R}^l}(a,b;t) = (1-t)a+tb$. Above, $\mathrm{err_i}$ is the error due to linear interpolation, and $\mathrm{err_b}$ is the base reconstruction error which occurs independently

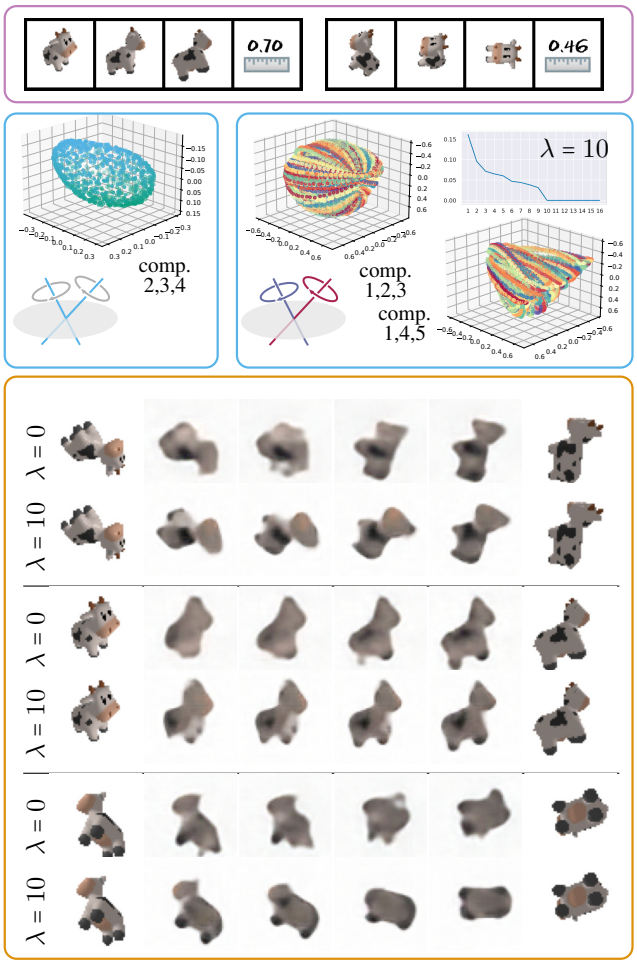

Figure 5. Results of our method for the rotated objects dataset (R). The middle boxes show projections of the obtained latent manifold $\phi(M)$: a submanifold for fixed rotation angle around all possible axes in S^2 (left) and all of $\phi(M)$ with rotations around the same axis in same color (right, once taking principal components 1, 2, 3, once 1, 4, 5). The additional curve illustrates the standard deviation along the principal components in latent space, indicating that 9 Euclidean dimensions are used for the embedding. The bottom box shows decoder outputs for linear interpolation in latent space between the codes of the first and last image. Such interpolation becomes feasible for higher bending weight λ, *even though the decoder was not trained for such codes.*

of interpolation. Figure 4 displays $\mathrm{err}(t)$ for different values of λ, showing a marked error reduction for increasing λ up to a saturation around $\lambda = 5$.

Additional dimensions exploited by the embedding. In our experiments, we set the latent space dimensionality l to commonly used values. In particular, we take l substantially larger than would minimally be required for a smooth embedding. This is reasonable since in applications the intrinsic dimensionality m is generally unknown and since this allows the encoder to improve on the flatness of the embedding at the expense of using more dimensions. For example, the flat torus should better be embedded as $S^1 \times S^1 \subset \mathbb{R}^4$ than as a torus in $\mathbb{R}^3$. Figure 5 shows that the encoder makes use of that freedom: even though $l = 5$ would be enough ($SO(3) \cong \mathbb{R}P^3$, which embeds into $\mathbb{R}^5$, but not $\mathbb{R}^4$ [10, 7]), the graph of the standard variation along the principal components of latent space shows that 9 Euclidean dimensions are used for the embedding. For the experiments with datasets (S) and (G), the embedding used 3 and 5 Euclidean dimensions, respectively.

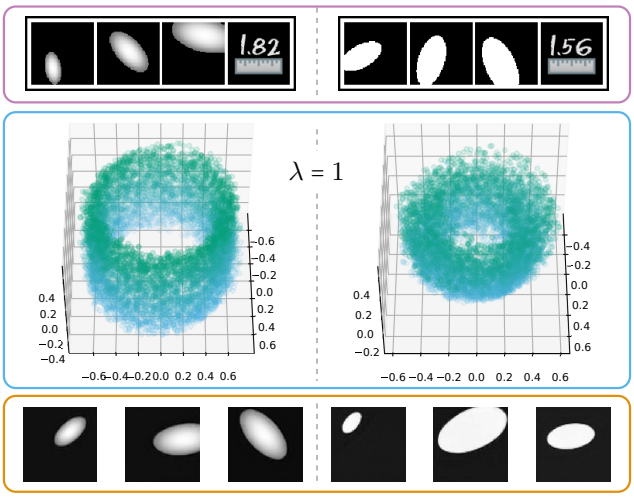

Figure 6. Obtained results for anisotropic Gaussian dataset (G); the shown latent manifold dimensions represent rotation (x-y-axis) and scale (z-axis). The noise in the right experiment stems from quantizing all input images to binary ones.

Noise in the embedding due to image quantization. To illustrate the regularization properties of our loss functional we avoided sources of noise in our experiments, as those would require additional tailored regularization. For the anisotropic Gaussian (G) we now illustrate what effect a simple type of noise can have on the embedding ϕ without additional regularization: we simply round all Gaussian images to binary images. This quantization makes ellipses in nearby positions, orientations and scalings harder to distinguish. Figure 6 right shows that a cylindrical structure of the resulting latent manifold $\phi(M)$ is still observable, though it is thickened and much less clean than on the left.

Acknowledgement This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via project 211504053 - Collaborative Research Center 1060 and via Germany's Excellence Strategy project 390685813 - Hausdorff Center for Mathematics and project 390685587 - Mathematics Münster: Dynamics-Geometry-Structure.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Andrei A. Agrachev, Davide Barilari, and Elisa Paoli. Volume geodesic distortion and Ricci curvature for Hamiltonian dynamics. *Annales de l'Institut Fourier*, 69(3):1187–1228, 2019.

[3] Matan Atzmon, Amos Gropp, and Yaron Lipman. Isometric autoencoders. *arXiv preprint arXiv:2006.09289*, 2020.

[4] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

[5] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[6] David L Donoho and Carrie Grimes. Image manifolds which are isometric to Euclidean space. *Journal of mathematical imaging and vision*, 23(1):5–24, 2005.

[7] W. Hantzsche. Einlagerung von Mannigfaltigkeiten in euklidische Räume. *Math. Z.*, 43(1):38–58, 1938.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[9] Emmanuel Hebey. *Sobolev spaces on Riemannian manifolds*, volume 1635. Springer Science & Business Media, 1996.

[10] Heinz Hopf. Systeme symmetrischer Bilinearformen und euklidische Modelle der projektiven Räume. *Vierteljschr. Naturforsch. Ges. Zürich*, 85(Beiblatt (Festschrift Rudolf Fueter)):165–177, 1940.

[11] Du Q. Huynh. Metrics for 3d rotations: Comparison and analysis. *J. Math. Imaging Vis.*, 35(2):155–164, Oct. 2009.

[12] Keizo Kato, Jing Zhou, Tomotake Sasaki, and Akira Nakagawa. Rate-distortion optimization guided autoencoder for isometric embedding in Euclidean latent space. In *International Conference on Machine Learning*, pages 5166–5176. PMLR, 2020.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14] Yunqian Ma and Yun Fu. *Manifold learning theory and applications*. CRC press, 2011.

[15] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[16] Alon Oring, Zohar Yakhini, and Yacov Hel-Or. Autoencoder image interpolation by shaping the latent space, 2020.

[17] Gautam Pai, Ronen Talmon, Alex Bronstein, and Ron Kimmel. Dimal: Deep isometric manifold learning using sparse geodesic sampling. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 819–828. IEEE, 2019.

[18] Erez Peterfreund, Ofir Lindenbaum, Felix Dietrich, Tom Bertalan, Matan Gavish, Ioannis G Kevrekidis, and Ronald R Coifman. Local conformal autoencoder for standardized data coordinates. *Proceedings of the National Academy of Sciences*, 117(49):30918–30927, 2020.

[19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[20] Marc Ranzato, Christopher Poultney, Sumit Chopra, Yann LeCun, et al. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19:1137, 2007.

[21] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[22] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Icml*, 2011.

[23] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[24] Ariel Schwartz and Ronen Talmon. Intrinsic isometric manifold learning with application to localization. *SIAM Journal on Imaging Sciences*, 12(3):1347–1391, 2019.

[25] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The Riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018.

[26] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[27] Alain Trouvé and Laurent Younes. Local geometry of deformable templates. *SIAM J. Math. Anal.*, 37(1):17–59, 2005.

[28] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[29] Laurent Younes. *Shapes and diffeomorphisms*, volume 171 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, 2010.